# Optimizing the automatic functional annotation of English intonation

*Daniel Hirst and Saandia Ali*

CNRS Laboratoire Parole et Langage
Université de Provence, Aix-en-Provence, France
{saandia.ali, daniel.hirst}@lpl-aix.fr

## Abstract

One of the fundamental aims of prosodic analysis is to provide a reliable means of extracting functional information (what prosody contributes to meaning) directly from prosodic form. It has been argued that an explicit model of the mapping from prosodic function to prosodic form could provide an objective way of approaching this task. In this presentation we look specifically at some of the problems of optimizing this mapping in order to extract the functional information automatically from the formal representation, hence ultimately directly from the acoustic data.

## 1. Introduction

Phonological models of intonation and prosody in general are notoriously theory dependent. Such models obviously need to address the problem of validation. What criteria can be used to decide that one particular representation of is more adequate than another.

There has, in the past, been little or no consensus on either the way to represent the forms of prosodic patterns, or on how to annotate prosodic functions, the way that speech prosody contributes to the interpretation of an utterance. In many systems of representation, formal and functional aspects are in fact conflated. This is the case for example with the most widely used annotation system, ToBI [8], originally designed as a standard for the representation of American English intonation, but later adapted to a number of other languages [7]. It has been argued [3] that maintaining a clear separation between prosodic form and prosodic function would provide a more effective means of validating or evaluating specific proposals.

In an earlier study [1], we suggested that a solution to the problem of evaluating models could be found by a process of analysis by synthesis - generating prosodic forms by an explicit mapping from a representation of prosodic functions and comparing the output with observable data.

We showed specifically how five successively more complex representations of English intonation patterns could be converted by rule into surface phonological representations of intonation using the INTSINT alphabet [5][3]. These can in turn be converted into phonetic representations by means of the Momel algorithm and the ouput can then be compared directly to the original recordings of 40 continuous 5 sentence passages from the Eurom1 corpus.

The five successive models we used to illustrate the process were :

a. ***none***. The intonation of the whole passage was modelled as a rising then falling pattern represented with INTSINT as [M-T…B] (where "-" indicates a fixed duration between the two tones and where [ and ] correspond to the beginning and end of the passage).

b. ***IU*** (Intonation Unit). Taking the boundaries of intonation units as functional annotation, each IU was modelled with the sequence [M-T…B] (where here [ and ] correspond to the beginning and end of the intonation unit).

c. ***terminal*** Intonation boundaries were marked as either terminal or non-terminal and modelled respectively as [M-T…B-B] and [M-T…B-H].

d. ***accent*** In addition to the tones assigned in model (c), in this model the T tone is aligned with the first accent and each subsequent accent is assigned a D tone.

e. ***nucleus*** In this model, a distinction was made between prenuclear accents coded as in model (d). and post-nuclear accents which are assigned a B tone.

f. ***emphasis*** in this model a distinction was made between non-emphatic patterns, coded as in model (e) and emphatic nuclei and heads, which were assigned more complex patterns, described in detail in [1].

This final pattern corresponds to the IF (intonation functions) annotation system originally proposed in [6] and more recently in [4]. Figure 1 shows an example of one of the Eurom1 passages with the functional information corresponding to model (f). In fact all the other models were derived from this functional representation by adapting the mapping rules appropriately.

**Figure 1.** *Passage fao30072 with the final functional annotation used in the analysis.*

# [Mu'nicipal *Fire 'Service 'speaking| # [We're 'trying to lo'cate an e'mergency *caller+ [who 'rang *off+ # [wi'thout 'giving any 'personal *details| # [He ap'peared to 'be on the 'local *network| # [He **c**on'nected on our !*line 'number+ # ['seven six *two+ 'five 'eight *four| # [!We'd ap'preciate im'mediate at'tempts to *trace him| [be'cause he 'sounded *desperate| #

By comparing the output of the successive models with the original recordings, it was shown that progressively increasing the complexity of the functional information leads to a corresponding increase in the correlation between the predicted values and the observed values, as can be seen in Table 1.

**Table 1** *mean and standard deviation of correlation coefficients for each Intonation Unit between the output of the model and the output from the hand-corrected Momel targets*

| model | none | IU | terminal | accent | emphasis |
|---|---|---|---|---|---|
| mean | 0.1238 | 0.5714 | 0.5215 | 0.5902 | 0.6573 |
| sd | 0.2929 | 0.2533 | 0.3386 | 0.2754 | 0.2238 |

We noted that while this result is encouraging, none of the mean correlations obtained were particularly good (although several individual values for specific sentences were much higher, often over 0.9). It should be noted, though, that in this preliminary study no attempt was made to optimise the parameters of the model. The mapping rules were formulated by hand, as was the specific functional annotation. Nevertheless, the study demonstrated that it is possible, using a technique of this type, to quantify the contribution of a particular functional annotation to the quality of the model.

In the rest of this paper we present some further research on this problem and in particular we make some preliminary steps towards deriving a functional representation such as that of Figure 1 directly from the acoustic signal rather than annotating it by hand as in our earlier presentation.

## 2. Optimizing the representation of prosodic form

The first stage of the optimization of the mapping between function and form was carried out using model (b) described above ie "IU" (Intonation Unit). The sentences recorded by four different speakers from the Eurom1 corpus were analyzed. The only functional information taken into account at this stage was the pause, hand-labelled and represented by "#". These pauses coincide with the five sentences read by each speaker within one sound file.

### 2.1. Optimizing the coding of INTSINT tones

It was decided to represent prosodic form between two pauses, using just four INTSINT tones: [t1-t2….t3-t4] where t1 and t2 were aligned respectively with the left and right boundaries and t3 and t4 were aligned at a fixed duration (200ms) after or before the initial or final boundary.
By means of a Praat script, and for each extract, all of the combinations of four INTSINT tones were tested except that the first tone of a unit was necessarily an absolute tone, either T, M or B, since a relative tone presupposes that there is a preceding target. For the other tones there were 8 possible tones at each point and we also included the possibility of no tones at all (_) at each point except the last. This gave a total of 1944 (=3*9*9*8) possible sequences of tones.

#### *2.1.1. Results of the experiment*

The F0 curve was calculated for each possible combination of tones and the correlation with the hand-corrected Momel curve was calculated. The combination of tones which obtained the best correlation on all of the extracts was then selected as shown in table 2.

### 2.2. Optimizing the alignment of the tones

Once an optimal sequence of tones was defined for each speaker: Fe, FF (MULL), Fa (TTLB), Fg (T_LB), we tried to optimize the alignment of these tones. Only the alignment tones t2 and t3 were optimized at this stage leaving t1 and t4 at a fixed offset.
Thus for speaker Fe the tones were fixed as follows:

- M fixed at an offset of 10ms from the left boundary
- U from 100 to 800 ms after the left boundary with steps of 50ms (15 iterations)
- L from 100 to 800 ms before the right boundary with steps of 50ms (15 iterations)
- L fixed at an offset of 10 ms from the right boundary

The pair of points obtaining the best correlation with the original Momel curve was selected using the same method as in the optimization of tones.

**Table 2.** *Best sequences of tones and target alignment after optimization*

| Speaker | Form | Left align t1 | Left align t2 | Right align t3 | Right align t4 | Mean Correlation |
|---|---|---|---|---|---|---|
| FA | T T L B | 0.01 | 0.2 | 0.7 | 0.01 | 0.666 |
| FE | M U L L | 0.01 | 0.2 | 0.7 | 0.01 | 0.683 |
| FF | M U L L | 0.01 | 0.2 | 0.4 | 0.01 | 0.650 |
| FG | T _ L B | 0.01 | 0.1 | 0.4 | 0.01 | 0.642 |

## 3. Optimizing the representation of prosodic function

The ultimate aim of this experiment is to extract the functional representation automatically from the formal representation. Hence, using the optimization of the sequence of tones and their alignment, our next attempt is to automatically detect intermediate boundaries of intonation units between two pauses.

### 3.1. Automatic segmentation of speech using melody only: from form to function

The minimal functional information used for this experiment was the word boundaries, on the fairly uncontroversial assumption that an Intonation Unit boundary must coincide with a word boundary.
A Praat script was used to test every possible location for an Intonation Unit boundary. In each case, the optimal sequence of tones, as determined from the first experiment, was applied

and the result of the correlation saved for comparison with other possible boundary locations.

The ultimate aim was to obtain a better correlation with the introduction of a new boundary and so the simple criterion for a new intonation unit boundary was that the correlation should be higher with two groups than with just one.

To begin with, the correlation was calculated for the whole passage, then each possible location was tested and the boundary corresponding to the highest correlation was retained provided that this correlation was higher than that obtained with no boundary. The process was then iterated until no further boundaries improved the correlation.

Two different series of detection of boundaries were run, the first using fixed parameters of tones and alignment as mentioned above and the second one allowing a variation of the alignment of the tones between the values giving the best correlations. Thus for Speaker Fe, the first experiment was run with the sequence of tones [M-U…L-L] and a fixed alignment of t2 and t3 (where t2=0.2 and t3=0.7) and in the second one t2 varied from 0.2 to 0.3 with three iterations and t3 from 0.65 to 0.8 with three iterations. It was expected that a model of detection allowing more flexibility in terms of target alignment would improve the detection of boundaries.

### *3.1.1. Evaluation of the detection of boundaries*

In order to evaluate, the automatic detection of boundaries, the results were compared with the annotation by the authors of this paper.(table 3**).** Boundaries were added by hand and qualified as terminal or non terminal. The results of the new experiments of boundary detections were classified in keeping with the type of experiment and the type of boundary (terminal/ non terminal).

**Table3.** *Results of the detection of boundaries. IU= hand-labeled Intonation Units. DB1= detection of boundaries with fixed alignment of t2-t3. DB2= variation of t2-t3's alignment. Nterm= Non terminal boundary/ Term= terminal boundary.*

| Boundtype | NB IU | NB DB1 | NB DB2 |
|---|---|---|---|
| Nterm before Pause | 95 | 95 | 95 |
| Term before Pause | 189 | 189 | 189 |
| NTerm | 328 | 48 | 70 |
| Next W after NTerm | | 7 | 6 |
| Next W0 after NTerm | | 35 | 43 |
| Term | 61 | 47 | 46 |
| Next W after Term | | 3 | 3 |
| Next W0 after Term | | 1 | 3 |
| (empty) | | 47 | 73 |
| Total | 673 | 472 | 528 |

### *3.1.2. Recurrent errors*

As can be seen in table 3, the second experiment of detection of boundaries (DB2) found 20 more boundaries than the first one. Most terminal boundaries were detected in both cases (46 out of 61/ 45 out of 61) while most non-terminal boundaries were missed (48 out of 328/ 70 out of 301). This suggests that a further optimisation of the model should be developed taking into account the distinction between terminal and non-terminal boundaries. Indeed, all of the optimised sequences of tones used for our speakers so far contain a final falling tone usually associated with terminal boundaries.

Furthermore, boundaries were frequently identified just one or two words after the one annotated by a linguist. (noted NextW/O in table 3). These were generally grammatical words like pronouns or articles, which are unaccented and sometimes also reduced. It seems that no major change of the shape of F0 can be noticed around these points so that a better correlation is obtained by adding them to the preceding intonation unit.

Examples for file Feo1074 :

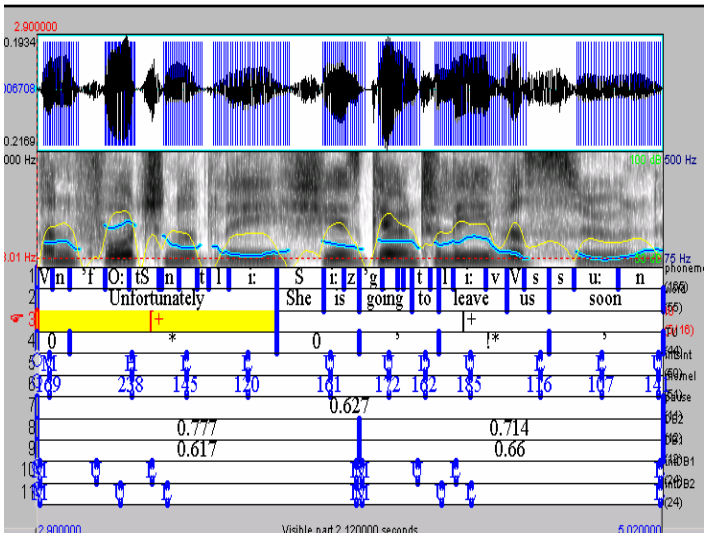
I have a problem | with my water softener.(Authors)

I have a problem with | my water softener. (DB1)

Unfortunately | she's going to leave us soon. (A)

Unfortunately she's | going to leave us soon. (DB1)

*Figure 2. Different models of detection of boundaries for the sentence : Unfortunately, she's going to leave us soon. , tier 3= IU, tier8=DB2, Tier 9= DB1*




There were many cases where no internal boundaries were added because the result of the correlation with two groups was only slightly inferior to the minimum required.
The requirements of the script may be too restrictive for some cases.

One last recurrent error is when the script takes a prominence for the beginning of a new intonation unit.

Examples for file Fer1094:

Maybe a salad | would be good. (A)

Maybe a | salad would be good. (DB1)

Mark it as top priority. (A)

Mark it as | top priority. (DB1)

In these examples a boundary was detected just before what was interpreted as the nucleus. The higher value of F0 at the beginning of the nucleus was interpreted as the beginning of a new IU. This suggests that a more complex model taking into account accents might provide a better fit.

### *3.1.3. Results*

In this section, we summarize the results obtained after the different steps of optimisation. After the automatic detection of intermediate boundaries, the correlations between the original Momel curves and the output of our model were evaluated anew.

**Table 4.** *mean and standard deviation of correlation coefficients for each Intonation Unit between the output of the model and the output from the hand-corrected Momel targets*

| | DB1 | | | DB2 | | | No bound | | |
|---|---|---|---|---|---|---|---|---|---|
| **correl** | **NB** | **Mean** | **SD** | **NB** | **Mean** | **SD** | **NB** | **Mean** | **SD** |
| >= 0.75 | 146 | 0.833 | 0.06 | 175 | 0.840 | 0.06 | 141 | 0.834 | 0.06 |
| >= 0.50 | 163 | 0.638 | 0.07 | 146 | 0.656 | 0.06 | 147 | 0.627 | 0.07 |
| < 0.50 | 44 | 0.197 | 0.40 | 31 | 0.176 | 0.42 | 65 | 0.251 | 0.34 |
| **Total** | **353** | **0.664** | **0.249** | **352** | **0.705** | **0.232** | **353** | **0.640** | **0.26** |

In both experiments (DB1 and DB2), the mean correlation for all speakers increased, rising from 0.57 to 0.66 and 0.705. (see table 1 and table 4). It must be added, though, that these results are not entirely comparable, since the first table sums up results taken from just one speaker as opposed to four speakers in the second. Furthermore, model (b) was based on the Intonation unit labelled by hand as opposed to an automatic detection in the second case.

## 4. Conclusions

In this paper, we make a first attempt at optimizing the parameters of a previously described model of the mapping between prosodic function and prosodic form. Initially, the only functional information taken into account was the pause, which, in read speech, can be fairly safely be assumed to correspond to a prosodic boundary [2]. The prosodic form was optimized as a sequence of four INTSINT tones the alignment of which was in turn optimized for four speakers. Our next endeavor was to use these optimal forms to try to automatically detect intermediate intonation unit boundaries. The results of this experiment were evaluated by comparison with the annotation of boundaries by hand. These were generally quite promising. A closer look at the errors and missing boundaries suggested the need to develop a more complex model which would take into account the type of boundary (terminal /non terminal) and prominences thus getting closer to the more complex models of synthesis described in section one.

## 5. References


[1] Ali, Saandia; Hirst, Daniel. 2007. Analysis by synthesis of English intonation patterns: generalising from form to function. in *Proceedings International Conference on Phonetic Sciences,* Saarbrücken, paper 1403.

[2] Cho, H & D.J. Hirst 2007. Empirical evidence for prosodic phrasing: pauses as linguistic annotation in Korean read speech. *Proceedings of Interspeech 2007*, Antwerp, Belgium.

[3] Hirst, D.J. 2005. Form and function in the representation of speech prosody. *Speech Communication* 46 (3-4), 334-347.

[4] Hirst, D.J. 2007. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. in *Proceedings International Conference on Phonetic Sciences,* Saarbrücken, paper 1443.

[5] Hirst, D.J., Di Cristo, A. & Espesser, R. 2000. Levels of representation and levels of analysis for intonation. in M. Horne (ed) *Prosody : Theory and Experiment*. Kluwer Academic Publishers, Dordrecht. 51-87.

[6] Hirst. D.J. 1977. *Intonative Features. A Syntactic Approach to English Intonation.* (Mouton, The Hague).

[7] Jun, Sun-Ah (ed.). 2006. *Prosodic Typology and Transcription: A Unified Approach.* Oxford University Press.

[8] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. 1992. ToBI : a Standard for Labelling English Prosody. *Proceedings ICSLP92* (2) 867- 870, Banff, Canada.